Open Journal of Statistics, **2012, 2, 60-72**
http://dx.doi.org/10.4236/ojs.2012.21007 Published Online January 2012 (http://www.SciRP.org/journal/ojs)

Comparative Analysis of Group Sequential Designs Tests for Randomized Controlled Clinical Trials: A Model Study on Two-Sided Tests for Comparing Two Treatments

Mehmet Ali Sungur, Emine Arzu Kanik
Department of Biostatistics and Medical Informatics, Medical Faculty, Mersin University, Mersin, Turkey
Email: malisungur@yahoo.com

Received July 13, 2011; revised August 18, 2011; accepted August 30, 2011

ABSTRACT

Clinical trials are usually long term studies and it seems impossible to reach all required subjects at the same time. Performing interim analyses and monitoring results may provide early termination of trial after obtaining significant results. The aim of this study is comparing group sequential tests in respect to advantage of sample size reduction and early termination. In this study, 4 test types used in group sequential designs were compared with fixed sample size design test and each other. Comparisons were done according to two-sided tests for comparing two treatments. In this sense, 1080 models were performed. In models, 2 different Type I errors, 2 different powers, 5 different analysis groups, 6 different effect sizes and 9 different variances selections were considered. All test types increased the maximum sample size in different manner, compared with fixed sample size design. Each test had different critical values to reject H_0 hypothesis, at the same type I error rate and number of analyses conditions. Selection of test type used in group sequential designs depends on a few characteristics, as reducing sample size, early termination and detecting minimal effect size. Test performance is highly related with selected Type I error rate, power and number of analyses. In addition to these statistical characteristics, researchers should decide test type with respect to other trial conditions as the issue of trial, reaching subjects easy or not and importance of early termination.

Keywords: Group Sequential Designs; Group Sequential Test Types; Interim Analysis; Monitoring

1. Introduction

Clinical trials are designed to detect differences between treatments with a certain power and Type I error rate. Investigators should ensure to design clinical trials that contain adequate statistical power and sample size. It takes a long time to reach required subject number at the same time. Data are accumulated periodically course of the trial. Thus, it may take a few years to enroll enough subjects to meet determined sample size at the beginning of the trial. Particularly, in clinical trials which have death risk or any potential harm this may be more difficult and reaching required sample size may cause more increasing of trial time. Therefore, it is an important interest of investigators substantially to analyze accumulated data in specific intervals and evaluate results. Performing interim analyses and monitoring results may provide early termination of trial after obtaining significant results corresponding superiority, inferiority or equivalency of new treatment according to standard method.

Clinical trials can be classified in two groups in term of sample size, as fixed sample size designs and sequential designs [1]. In fixed sample designs, sample size is calculated at the beginning of the trial, and data are analyzed once after all required subjects enrolled. In sequential designs, sample size is calculated at the beginning of the trial similarly, but data are analyzed periodically by interim analyses as the trial going on and a final analysis is done at the finishing of the trial if required. Results of each interim analysis are evaluated to decide stopping or continuing the trial, and thereby the trial is monitoring [2-4].

Sequential designs were initially developed for economical reasons. Early termination for a trial that have positive result means that a new product can be used sooner. If the trial have negative results, early termination ensures saving from sources. Sequential designs typically serve to savings in sample size, time and cost of the trial comparing with fixed sample size designs [5-7].

There are several reasons for monitoring a trial, and decide to stopping or continuing it. In medical researches possible side effects, quality of life, cost or availability of alternative treatments can not be known at the beginning of the trial [5]. The most important reason is stop treating subjects with an ineffective treatment, when results show

that test treatment is superior, inferior or equivalent to the standard treatment.

Sequential designs are categorized in three groups: fully sequential designs, group sequential designs and flexible sequential designs [1,5]. In group sequential designs, interim analyses are done periodically at certain times determined at the beginning of the trial. Group sequential designs require determination of number and time of interim analyses at the beginning of the trial and remain constant. Interim analyses must be done by equal intervals [1].

Group sequential designs based on the evaluation of results obtained interim analysis of data collected from each patient group with predetermined sample size [1]. There are many statistical criteria that controlled Type I error rate during periodic analyses. At each interim analysis test value calculated and compared with critical value of test. These critical values vary according to number of interim analyses and Type I error rate selected. Commonly used group sequential tests are suggested by Pocock and O'Brien & Fleming [6,7]. They have been improved for common test statistics used to compare means, medians, proportions or survival curves. Group sequential designs required determination of number and time of interim analyses at the beginning of the trial and remain constant. Interim analyses must be done in equal intervals [1].

The aim of this study is to evaluate four types (Pocock, O'Brien & Fleming, Wang & Tsiatis and Haybittle-Peto tests) of group sequential designs' tests used to compare means of two treatments comparatively. The comparisons were done in respect to advantage of sample size reduction, potential of early termination and detecting minimum differences between treatments at the same conditions.

2. Material and Methods

2.1. Two-Sided Tests for Comparing Two Treatments

In two sided hypothesis tests, the null hypothesis (H_0) referring "there is no statistically significant difference between two treatments" is controlled against the alternative hypothesis (H_1) referring "there is a statistically significant difference between two treatments".

When treatments' means distributing normally with a known variance, the test statistic calculating is Z. In fixed sample size designs, when the value of Z statistic calculated is equal to or larger than a c value named "critical value" the null hypothesis (H_0) is rejected while it is accepted when the value of Z statistic is less. Determination of critical value based on the Type I error rate selected. Type I error is usually determined as 0.05 while 0.01 or 0.001 values are selected when the study has death risk, irreversible harms or potential risks. The formula of Z statistic to compare means of two treatments as A and B distributing normally with a known variance, and including n subjects is as follow [8,9]:

$$Z = \frac{\left(\bar{X}_A - \bar{X}_B\right) - \left(\mu_A - \mu_B\right)}{\sqrt{\frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B}}} \qquad (2.1.1)$$

Required sample size in each treatment group for comparing two independent groups is calculated as following way [5,8-11]:

$$n_f = \frac{\left(Z_{\alpha/2} + Z_\beta\right)^2 \cdot \left(\sigma_A^2 + \sigma_B^2\right)}{\left(\mu_A - \mu_B\right)^2} \qquad (2.1.2)$$

2.2. Group Sequential Designs

In group sequential designs, number of analysis (K) and required sample size (m) in each group for each analysis is determined initially while two treatments are comparing. Total number of analyses in a group sequential design is K, consisting of $K - 1$ interim analyses and a final analysis. The maximum subject number enroll to study is $2mK$. The formula of Z_k statistic to compare two means if A and B distributing normally with a known variance, and including $2m$ subjects is as follow [5,9]:

$$Z_k = \frac{\left(\bar{X}_A - \bar{X}_B\right) - \left(\mu_A - \mu_B\right)}{\sqrt{\frac{\sigma_A^2}{m_A} + \frac{\sigma_B^2}{m_B}}},\ k = 1,\ 2, \cdots,\ K \qquad (2.2.1)$$

Maximum sample sizes for each test types are different and calculated by multiplying n_f in Equation (2.1.2) with a special factor R varying for each test type and each number of analysis.

In group sequential designs, since the number of statistical analyses is more than one, to protect total Type I error rate (α), Type I error rate (α_k) for each analysis should be determined by allocating the total Type I error rate (α) to each analysis. And the c_k critical values are determined according to these Type I error rates (α_k) [1-3, 5,12]. Interim analyses are done after collecting data from each $2m$ subjects groups periodically and Z_k test statistic is calculated for each analysis. When the value of Z_k statistic calculated is equal to or larger than c_k critical value the null hypothesis (H_0) is rejected and the trial is terminated, referring as "positive result". If the value of Z_k statistic calculated is less than c_k critical value, the trial continuing by adding a new $2m$ subjects group. If there is no positive result until final analysis and if still $Z_k < c_k$ at the final analysis, than trial is terminated accepting the null hypothesis, referring as "negative result" [1,3, 5,9].

2.2.1. Pocock Test

The value of *Z* statistic after each interim analyses and final analysis is calculated with the formula given at Equation (2.2.1). The critical values of Pocock test that compared with the value of *Z* statistics are denoted as $C_P(K,\alpha)$. The critical values $C_P(K,\alpha)$ varying based on total Type I error rate (α) and number of analyses (K), and remain constant for all interim analyses and final analysis [5,9].

The maximum sample size for Pocock test is calculated by multiplying n_f in Equation (2.1.2) with $R_P(K,\alpha,\beta)$ values. The $R_P(K,\alpha,\beta)$ values varying based on total Type I error rate (α), Type II error rate (β) and number of analyses (K). Subject number for each treatment in each interim analysis is calculated as follow [5,9]:

$$m = \left(R_P(K,\ \alpha,\ \beta)\cdot n_f\right)\Big/K \qquad (2.2.2)$$

2.2.2. O'Brien & Fleming Test

The value of *Z* statistic after each interim analyses and final analysis is calculated in the same way with the formula given at Equation (2.2.1). The critical values of O'Brien & Fleming test that compared with the value of *Z* statistics are denoted as $C_B(K,\ \alpha)$ for final analysis. For interim analyses, the critical values are obtained by multiplying $C_B(K,\ \alpha)$ with $\sqrt{K/k}$. The critical values $C_B(K,\ \alpha)$ varying based on total Type I error rate (α) and number of analyses (K), and are different for each interim analyses and final analysis [5,9]:

The maximum sample size for O'Brien & Fleming test is calculated by multiplying n_f in Equation (2.1.2) with $R_B(K,\alpha,\beta)$ values. The $R_B(K,\alpha,\beta)$ values varying based on total Type I error rate (α), Type II error rate (β) and number of analyses (K). Subject number for each treatment in each interim analysis is calculated as follow [5,9]:

$$m = \left(R_B(K,\ \alpha,\ \beta)\cdot n_f\right)\Big/K \qquad (2.2.3)$$

2.2.3. Wang & Tsiatis Test

There is a Δ parameter for Wang & Tsiatis test differently from other tests and certain values of this parameter makes Wang & Tsiatis test the same with Pocock and O'Brien & Fleming tests. Wang & Tsiatis test is same with Pocock test when $\Delta = 0.50$ and with O'Brien & Fleming test when $\Delta = 0$. Values of Δ between 0 - 0.5 gives critical values between Pocock and O'Brien & Fleming tests. Also, the value of *Z* statistic after each interim analyses and final analysis is calculated with the formula given at Equation (2.2.1). The critical values of Wang & Tsiatis test that compared with the value of *Z* statistics are denoted as $C_{WT}(K,\ \alpha,\ \Delta)$ for final analysis. For interim analyses, the critical values are obtained by multiplying $C_{WT}(K,\ \alpha,\ \Delta)$ with $(k/K)^{\Delta-1/2}$. The critical values $C_{WT}(K,\ \alpha,\ \Delta)$ varying based on total Type I error rate (α) and number of analyses (K), and are different for each interim analyses and final analysis [5, 9].

The maximum sample size for Wang & Tsiatis test is calculated in the same manner by multiplying n_f in Equation (2.1.2) with $R_{WT}(K,\alpha,\beta,\Delta)$ values. The $R_{WT}(K,\alpha,\beta,\Delta)$ values vary based on total Type I error rate (α), Type II error rate (β), number of analyses (K) and value of Δ. Subject number for each treatment in each interim analysis is calculated as follow [5,9].

$$m = \left(R_{WT}(K,\alpha,\beta,\Delta)\cdot n_f\right)\Big/K \qquad (2.2.4)$$

2.2.4. Haybittle-Peto Test

Calculation of *Z* statistic after each interim analyses and final analysis is the same with other tests, with the formula given at Equation (2.2.1). Haybittle-Peto test suggested that, in $k < K$ analyses, namely in all interim analyses, H_0 can be rejected only if $|Z_k| \geq 3$. So, critical values of this test denoted as $C_{HP}(K,\ \alpha)$ are constant ($c_1 = \cdots = c_{K-1} = 3$) for all interim analyses. It is different only for final analysis The critical values $C_{HP}(K,\ \alpha)$ for final analysis is varying based on total Type I error rate (α) and number of analyses (K) [5,9].

The maximum sample size for Haybittle-Peto test is calculated by multiplying n_f in Equation (2.1.2) with $R_{HP}(K,\alpha,\beta)$ values. The $R_{HP}(K,\alpha,\beta)$ values vary based on total Type I error rate (α), Type II error rate (β) and number of analyses (K). Subject number for each treatment in each interim analysis is calculated as follow [5,9].

$$m = \left(R_{HP}(K,\alpha,\beta)\cdot n_f\right)\Big/K \qquad (2.2.5)$$

2.3. Models

In this study, four test types (Pocock, O'Brien & Fleming, Wang & Tsiatis and Haybittle-Peto tests) used to test difference between to treatment in group sequential designs were compared with fixed sample size design test and each other. In this sense, 10080 models were performed. In models, 2 different Type I errors (α), 2 different powers (1-β), 14 different number of analyses (K), 15 different effect sizes (d) and 12 different variances (σ^2) selections were considered:

$$\alpha = 0.05 \text{ and } 0.01$$

$$(1-\beta) = 0.90 \text{ and } 0.80$$

$$K = 1;\ 2;\ 3;\ 4;\ 5;\ 6;\ 7;\ 8;\ 9;\ 10;\ 11;\ 12;\ 15 \text{ and } 20$$

$$d = 0,5;\ 1,0;\ 1,5;\ 2,0;\ 2,5;\ 3,0;\ 3,5;\ 4,0;\ 4,5;\ 5,0;$$

$$d = (\text{continue})\ 6,0;\ 7,0;\ 8,0;\ 9,0 \text{ and } 10,0$$

$$\sigma^2 = 1;\ 2;\ 3;\ 4;\ 5;\ 6;\ 7;\ 8;\ 9;\ 10;\ 15 \text{ and } 20$$

Sample size calculations for large effect sizes were too small and critical values according to analysis group numbers can be calculated by iteration [5,13], so 1080 of these 10080 models were used. In these models, 2 different Type I errors (α), 2 different powers (1-β), 5 different number of analyses (K), 6 different effect sizes (d) and 9 different variances (σ^2) selections were considered:

$$\alpha = 0.05 \text{ and } 0.01$$

$$(1-\beta) = 0.90 \text{ and } 0.80$$

$$K = 1;\ 5;\ 10;\ 15 \text{ and } 20$$

$$d = 0.5;\ 1.0;\ 1.5;\ 2.0;\ 2.5 \text{ and } 3.0$$

$$\sigma^2 = 1;\ 2;\ 3;\ 4;\ 5;\ 6;\ 7;\ 8 \text{ and } 9$$

Critical values to reject null hypothesis (H_0) and maximum sample sizes required for all test types were determined for each combination. In each combination, these four test types were compared with each other and fixed sample size design test, and advantages and disadvantages of tests were examined in same conditions. Haybittle-Peto and Wang & Tsiatis tests can only be used for $\alpha = 0.05$ Type I error rate [5,13]. So, these tests were only compared for $\alpha = 0.05$ combinations. In other combinations, only Pocock and O'Brien & Fleming test were compared with each other and fixed sample size design test. In addition, critical values of all test types decreasing from first interim analysis to final analysis. So, critical values for different number of analysis not included in tables can be calculated by iteration [5,13]. Results are summarized with tables.

3. Results

Critical values of each test types to reject null hypothesis (H_0) at $\alpha = 0.05$ level were shown at **Tables 1-4**. Also, sample size calculations for each test types to detect 3 different effect sizes with 6 several variances at 2 different $1-\beta$ level and $\alpha = 0.05$ level were shown at **Tables 5-10**. Since Haybittle-Peto and Wang & Tsiatis tests can only be used at $\alpha = 0.05$ level, comparisons at $\alpha = 0.01$ level were done only for Pocock and O'Brien & Fleming tests. Advantages and disadvantage of these tests according to each other at $\alpha = 0.01$ level were same with the condition at $\alpha = 0.05$ level. So, results of $\alpha = 0.01$ level combinations were not shown. In addition, sample sizes calculated for larger effect sizes were very small as 3.9 and some of them not possible in practice. Thereby some of these sample sizes were similar and not comparable. So, sample sizes for $d = 2.0$, $d = 2.5$ and $d = 3.0$ were not shown. Sample sizes calculated according to different variance selections increasing as variances increase and can be calculated by iteration. So, sample sizes for $\sigma^2 = 4$, $\sigma^2 = 6$, and $\sigma^2 = 8$ were not shown.

Critical value of fixed sample size design test to reject null hypothesis (H_0) at $\alpha = 0.05$ level is $c_K = 1.96$. The nearest critical value of group sequential tests at final analysis was obtained for Haybittle-Peto test and the furthest one was obtained for Pocock test. And it did not change according to the number of analysis (**Tables 1-4**).

When $K = 5$, Pocock test had the lowest critical values that might detect smaller effect sizes in first three interim analyses while O'Brien & Fleming test had the lowest critical values that might detect smaller effect sizes at 4th interim analysis and final analysis. This changing was observed at 7th interim analysis when $K = 10$, at 10th interim analysis when $K = 15$ and at 13th interim analysis when $K = 20$. Wang & Tsiatis test was always placed between these two tests. It was closed to O'Brien & Fleming test for small Δ values and started to close up to Pocock with increasing Δ values. Critical values of Wang & Tsiatis test for early interim analyses were getting closer to Pocock test as the Δ values increase, and they were decreasing for later interim analyses. Similarly, critical values of Wang & Tsiatis test were getting closer to O'Brien & Fleming test as the Δ values decrease, and they were decreasing for later interim analyses in parallel with critical values of O'Brien & Fleming test. When $K = 5$, Pocock test had lower critical values that might detect smaller effect sizes than Wang & Tsiatis test for all Δ values in first two or three interim analyses, while Wang & Tsiatis test had lower critical values that might detect smaller effect sizes in last two interim analyses and final analysis. This changing was observed at 5th - 7th interim analysis when $K = 10$, at 7th - 10th interim analysis when $K = 15$ and at 8th - 11th interim analysis when $K = 20$. Number of analysis which this changing was observed, was varying according to value of Δ. Haybittle-Peto test had a different way as having a constant critical value for all interim analysis. This critical value was placed between critical values of Pocock and O'Brien & Fleming tests for early interim analyses, and became higher from them after a few interim analysis. This changing was observed at 3rd interim analysis when $K = 5$, at 5th interim analysis when $K = 10$, at 8th interim analysis when $K = 15$ and at 11th interim analysis when $K = 20$. Only for final analysis Haybittle-Peto test had a different critical value that the nearest one to fixed sample size design test (**Tables 1-4**).

All of the group sequential tests were required more sample size than the fixed sample size designs (**Tables 5-10**). This increase was depending on effect size. Difference between sample sizes for group sequential tests and fixed sample size design was minimal when the effect size was large. Even they were similar beginning from $d = 2.0$.

Table 1. Critical values for α = 0.05 and K = 5.

K	C_P	C_B	$C_{WT}, \Delta = 0.10$	$C_{WT}, \Delta = 0.25$	$C_{WT}, \Delta = 0.40$	C_{HP}
1	2.413	4.562	3.937	3.194	2.663	3.000
2	2.413	3.226	2.984	2.686	2.485	3.000
3	2.413	2.634	2.357	2.427	2.386	3.000
4	2.413	2.281	2.261	2.259	2.318	3.000
5	2.413	2.040	2.068	2.136	2.267	1.990

K: Number of analysis, *C*: Critical values of tests; C_P: Pocock; C_B: O'Brien & Fleming; C_{WT}: Wang & Tsiatis; C_{HP}: Haybittle-Peto.

Table 2. Critical values for α = 0.05 and K = 10.

K	C_P	C_B	$C_{WT}, \Delta = 0.10$	$C_{WT}, \Delta = 0.25$	$C_{WT}, \Delta = 0.40$	C_{HP}
1	2.555	6.600	5.325	3.910	2.965	3.000
2	2.555	4.667	4.036	3.288	2.766	3.000
3	2.555	3.810	3.432	2.971	2.656	3.000
4	2.555	3.300	3.059	2.765	2.581	3.000
5	2.555	2.951	2.797	2.615	2.524	3.000
6	2.555	2.694	2.601	2.499	2.478	3.000
7	2.555	2.494	2.445	2.404	2.441	3.000
8	2.555	2.333	2.318	2.325	2.408	3.000
9	2.555	2.200	2.211	2.258	2.380	3.000
10	2.555	2.087	2.120	2.199	2.355	2.021

K: Number of analysis, *C*: Critical values of tests; C_P: Pocock; C_B: O'Brien & Fleming; C_{WT}: Wang & Tsiatis; C_{HP}: Haybittle-Peto.

Table 3. Critical values for α = 0.05 and K = 15.

K	C_P	C_B	$C_{WT}, \Delta = 0.10$	$C_{WT}, \Delta = 0.25$	$C_{WT}, \Delta = 0.40$	C_{HP}
1	2.626	8.172	6.340	4.387	3.143	3.000
2	2.626	5.778	4.805	3.689	2.932	3.000
3	2.626	4.718	4.085	3.333	2.816	3.000
4	2.626	4.086	3.641	3.102	2.736	3.000
5	2.626	3.655	3.330	2.934	2.675	3.000
6	2.626	3.336	3.096	2.803	2.627	3.000
7	2.626	3.089	2.911	2.697	2.587	3.000
8	2.626	2.889	2.760	2.608	2.553	3.000
9	2.626	2.724	2.633	2.533	2.523	3.000
10	2.626	2.584	2.524	2.467	2.496	3.000
11	2.626	2.464	2.429	2.409	2.473	3.000
12	2.626	2.359	2.346	2.357	2.451	3.000
13	2.626	2.267	2.272	2.310	2.432	3.000
14	2.626	2.184	2.206	2.268	2.414	3.000
15	2.626	2.110	2.146	2.229	2.397	2.046

K: Number of analysis, *C*: Critical values of tests; C_P: Pocock; C_B: O'Brien & Fleming; C_{WT}: Wang & Tsiatis; C_{HP}: Haybittle-Peto.

Table 4. Critical values for $\alpha = 0.05$ and $K = 20$.

K	C_P	C_B	$C_{WT}, \Delta = 0.10$	$C_{WT}, \Delta = 0.25$	$C_{WT}, \Delta = 0.40$	C_{HP}
1	2.672	9.508	7.166	4.754	3.269	3.000
2	2.672	6.723	5.431	3.998	3.050	3.000
3	2.672	5.489	4.618	3.612	2.929	3.000
4	2.672	4.754	4.116	3.362	2.846	3.000
5	2.672	4.252	3.764	3.179	2.783	3.000
6	2.672	3.882	3.500	3.038	2.733	3.000
7	2.672	3.594	3.290	2.923	2.691	3.000
8	2.672	3.362	3.119	2.827	2.656	3.000
9	2.672	3.169	2.976	2.745	2.624	3.000
10	2.672	3.007	2.853	2.673	2.597	3.000
11	2.672	2.867	2.746	2.610	2.572	3.000
12	2.672	2.745	2.652	2.554	2.550	3.000
13	2.672	2.637	2.569	2.504	2.530	3.000
14	2.672	2.541	2.494	2.458	2.511	3.000
15	2.672	2.455	2.426	2.416	2.494	3.000
16	2.672	2.377	2.364	2.377	2.478	3.000
17	2.672	2.306	2.307	2.341	2.463	3.000
18	2.672	2.241	2.255	2.308	2.449	3.000
19	2.672	2.181	2.207	2.277	2.435	3.000
20	2.672	2.126	2.162	2.248	2.423	2.068

K: Number of analysis, C: Critical values of tests; C_P: Pocock; C_B: O'Brien & Fleming; C_{WT}: Wang & Tsiatis; C_{HP}: Haybittle-Peto.

Table 5. Sample sizes for $\alpha = 0.05$, $(1 - \beta) = 0.90$, and $d = 0.5$.

K	n	$\sigma^2 = 1$	$\sigma^2 = 2$	$\sigma^2 = 3$	$\sigma^2 = 5$	$\sigma^2 = 7$	$\sigma^2 = 9$
1	n_f	84.0	168.0	251.9	419.9	587.9	755.8
5	n_P	101.4	202.7	304.1	506.8	709.6	912.3
	n_B	86.2	172.3	258.5	430.8	603.2	775.5
	$n_{WT}, \Delta = 0.10$	87.1	174.2	261.3	435.4	609.6	783.8
	$n_{WT}, \Delta = 0.25$	89.5	179.0	268.6	447.6	626.7	805.7
	$n_{WT}, \Delta = 0.40$	94.8	189.6	284.4	474.1	663.7	853.3
	n_{HP}	85.2	170.3	255.5	425.8	596.1	766.4
10	n_P	106.7	213.5	320.2	533.7	747.2	960.7
	n_B	87.1	174.2	261.3	435.4	609.6	783.8
	$n_{WT}, \Delta = 0.10$	88.2	176.4	264.5	440.9	617.3	793.6
	$n_{WT}, \Delta = 0.25$	91.0	181.9	272.9	454.8	636.7	818.6
	$n_{WT}, \Delta = 0.40$	97.3	194.7	292.0	486.7	681.3	876.0
	n_{HP}	86.5	173.0	259.5	432.5	605.5	778.5
15	n_P	109.6	219.2	328.8	548.0	767.2	986.4
	n_B	87.5	175.0	262.5	437.5	612.6	787.6
	$n_{WT}, \Delta = 0.10$	88.6	177.2	265.8	443.0	620.2	797.4
	$n_{WT}, \Delta = 0.25$	91.5	183.1	274.6	457.7	640.8	823.9
	$n_{WT}, \Delta = 0.40$	98.4	196.9	295.3	492.1	689.0	885.8
	n_{HP}	87.6	175.2	262.8	438.0	613.1	788.3
20	n_P	111.4	222.9	334.3	557.2	780.1	1003.0
	n_B	87.8	175.5	263.3	438.8	614.3	789.8
	$n_{WT}, \Delta = 0.10$	88.9	177.7	266.6	444.3	622.0	799.7
	$n_{WT}, \Delta = 0.25$	91.9	183.7	275.6	459.4	643.1	826.9
	$n_{WT}, \Delta = 0.40$	99.1	198.2	297.3	495.5	693.7	891.9
	n_{HP}	88.6	177.2	265.8	443.0	620.2	797.4

K: Number of analysis, n: Sample sizes for tests; n_P: Pocock; n_B: O'Brien & Fleming; n_{WT}: Wang & Tsiatis; n_{HP}: Haybittle-Peto.

Table 6. Sample sizes for α = 0.05, (1 – β) = 0.90, and d = 1.0.

K	n	$\sigma^2 = 1$	$\sigma^2 = 2$	$\sigma^2 = 3$	$\sigma^2 = 5$	$\sigma^2 = 7$	$\sigma^2 = 9$
1	n_f	21.0	42.0	63.0	105.0	147.0	189.0
5	n_P	25.3	50.7	76.0	126.7	177.4	228.1
	n_B	21.5	43.1	64.6	107.7	150.8	193.9
	$n_{WT}, \Delta = 0.10$	21.8	43.5	65.3	108.9	152.4	195.9
	$n_{WT}, \Delta = 0.25$	22.4	44.8	67.1	111.9	156.7	201.4
	$n_{WT}, \Delta = 0.40$	23.7	47.4	71.1	118.5	165.9	213.3
	n_{HP}	21.3	42.6	63.9	106.4	149.0	191.6
10	n_P	26.7	53.4	80.1	133.4	186.8	240.2
	n_B	21.8	43.5	65.3	108.9	152.4	195.9
	$n_{WT}, \Delta = 0.10$	22.0	44.1	66.1	110.2	154.3	198.4
	$n_{WT}, \Delta = 0.25$	22.7	45.5	68.2	113.7	159.2	204.6
	$n_{WT}, \Delta = 0.40$	24.3	48.7	73.0	121.7	170.3	219.0
	n_{HP}	21.6	43.3	64.9	108.1	151.4	194.6
15	n_P	27.4	54.8	82.2	137.0	191.8	246.6
	n_B	21.9	43.8	65.6	109.4	153.1	196.9
	$n_{WT}, \Delta = 0.10$	22.1	44.3	66.4	110.7	155.0	199.3
	$n_{WT}, \Delta = 0.25$	22.9	45.8	68.7	114.4	160.2	206.0
	$n_{WT}, \Delta = 0.40$	24.6	49.2	73.8	123.0	172.2	221.5
	n_{HP}	21.9	43.8	65.7	109.5	153.3	197.1
20	n_P	27.9	55.7	83.6	139.3	195.0	250.7
	n_B	21.9	43.9	65.8	109.7	153.6	197.5
	$n_{WT}, \Delta = 0.10$	22.2	44.4	66.6	111.1	155.5	199.9
	$n_{WT}, \Delta = 0.25$	23.0	45.9	68.9	114.8	160.8	206.7
	$n_{WT}, \Delta = 0.40$	24.8	49.5	74.3	123.9	173.4	223.0
	n_{HP}	22.1	44.3	66.4	110.7	155.0	199.3

K: Number of analysis, n: Sample sizes for tests; n_P: Pocock; n_B: O'Brien & Fleming; n_{WT}: Wang & Tsiatis; n_{HP}: Haybittle-Peto.

Table 7. Sample sizes for α = 0.05, (1 – β) = 0.90, and d = 1.5.

K	n	$\sigma^2 = 1$	$\sigma^2 = 2$	$\sigma^2 = 3$	$\sigma^2 = 5$	$\sigma^2 = 7$	$\sigma^2 = 9$
1	n_f	9.3	18.7	28.0	46.7	65.3	84.0
5	n_P	11.3	22.5	33.8	56.3	78.8	101.4
	n_B	9.6	19.1	28.7	47.9	67.0	86.2
	$n_{WT}, \Delta = 0.10$	9.7	19.4	29.0	48.4	67.7	87.1
	$n_{WT}, \Delta = 0.25$	9.9	19.9	29.8	49.7	69.6	89.5
	$n_{WT}, \Delta = 0.40$	10.5	21.1	31.6	52.7	73.7	94.8
	n_{HP}	9.5	18.9	28.4	47.3	66.2	85.2
10	n_P	11.9	23.7	35.6	59.3	83.0	106.7
	n_B	9.7	19.4	29.0	48.4	67.7	87.1
	$n_{WT}, \Delta = 0.10$	9.8	19.6	29.4	49.0	68.6	88.2
	$n_{WT}, \Delta = 0.25$	10.1	20.2	30.3	50.5	70.7	91.0
	$n_{WT}, \Delta = 0.40$	10.8	21.6	32.4	54.1	75.7	97.3
	n_{HP}	9.6	19.2	28.8	48.1	67.3	86.5
15	n_P	12.2	24.4	36.5	60.9	85.2	109.6
	n_B	9.7	19.4	29.2	48.6	68.1	87.5
	$n_{WT}, \Delta = 0.10$	9.8	19.7	29.5	49.2	68.9	88.6
	$n_{WT}, \Delta = 0.25$	10.2	20.3	30.5	50.9	71.2	91.5
	$n_{WT}, \Delta = 0.40$	10.9	21.9	32.8	54.7	76.6	98.4
	n_{HP}	9.7	19.5	29.2	48.7	68.1	87.6
20	n_P	12.4	24.8	37.1	61.9	86.7	111.4
	n_B	9.8	19.5	29.3	48.8	68.3	87.8
	$n_{WT}, \Delta = 0.10$	9.9	19.7	29.6	49.4	69.1	88.9
	$n_{WT}, \Delta = 0.25$	10.2	20.4	30.6	51.0	71.5	91.9
	$n_{WT}, \Delta = 0.40$	11.0	22.0	33.0	55.1	77.1	99.1
	n_{HP}	9.8	19.7	29.5	49.2	68.9	88.6

K: Number of analysis, n: Sample sizes for tests; n_P: Pocock; n_B: O'Brien & Fleming; n_{WT}: Wang & Tsiatis; n_{HP}: Haybittle-Peto.

Table 8. Sample sizes for $\alpha = 0.05$, $(1 - \beta) = 0.80$, and $d = 0.5$.

K	n	$\sigma^2 = 1$	$\sigma^2 = 2$	$\sigma^2 = 3$	$\sigma^2 = 5$	$\sigma^2 = 7$	$\sigma^2 = 9$
1	n_f	62.7	125.4	188.2	313.6	439.0	564.5
5	n_P	77.1	154.2	231.2	385.4	539.6	693.7
	n_B	64.5	129.0	193.4	322.4	451.3	580.3
	$n_{WT}, \Delta = 0.10$	65.2	130.5	195.7	326.1	456.6	587.1
	$n_{WT}, \Delta = 0.25$	67.2	134.5	201.7	336.2	470.7	605.1
	$n_{WT}, \Delta = 0.40$	71.6	143.3	214.9	358.1	501.4	644.6
	n_{HP}	63.7	127.3	191.0	318.3	445.6	572.9
10	n_P	81.6	163.2	244.8	408.0	571.2	734.4
	n_B	65.2	130.5	195.7	326.1	456.6	587.1
	$n_{WT}, \Delta = 0.10$	66.1	132.2	198.3	330.5	462.7	595.0
	$n_{WT}, \Delta = 0.25$	68.3	136.6	204.9	341.5	478.1	614.7
	$n_{WT}, \Delta = 0.40$	73.7	147.4	221.1	368.5	515.9	663.3
	n_{HP}	64.8	129.6	194.4	323.9	453.5	583.1
15	n_P	83.9	167.8	251.8	419.6	587.4	755.3
	n_B	65.5	131.1	196.6	327.7	458.8	589.9
	$n_{WT}, \Delta = 0.10$	66.4	132.8	199.3	332.1	464.9	597.8
	$n_{WT}, \Delta = 0.25$	68.8	137.6	206.4	344.0	481.6	619.2
	$n_{WT}, \Delta = 0.40$	74.6	149.1	223.7	372.9	522.0	671.2
	n_{HP}	65.7	131.5	197.2	328.7	460.1	591.6
20	n_P	85.5	171.0	256.5	427.4	598.4	769.4
	n_B	65.7	131.3	197.0	328.3	459.7	591.0
	$n_{WT}, \Delta = 0.10$	66.6	133.2	199.8	333.0	466.3	599.5
	$n_{WT}, \Delta = 0.25$	69.1	138.1	207.2	345.3	483.4	621.5
	$n_{WT}, \Delta = 0.40$	75.1	150.2	225.2	375.4	525.5	675.7
	n_{HP}	66.5	133.1	199.6	332.7	465.8	598.9

K: Number of analysis, n: Sample sizes for tests; n_P: Pocock; n_B: O'Brien & Fleming; n_{WT}: Wang & Tsiatis; n_{HP}: Haybittle-Peto.

Table 9. Sample sizes for α = 0.05, (1 – β) = 0.80, and d = 1.0.

K	n	$\sigma^2 = 1$	$\sigma^2 = 2$	$\sigma^2 = 3$	$\sigma^2 = 5$	$\sigma^2 = 7$	$\sigma^2 = 9$
1	n_f	15.7	31.4	47.0	78.4	109.8	141.1
5	n_P	19.3	38.5	57.8	96.4	134.9	173.4
	n_B	16.1	32.2	48.4	80.6	112.8	145.1
	$n_{WT}, \Delta = 0.10$	16.3	32.6	48.9	81.5	114.2	146.8
	$n_{WT}, \Delta = 0.25$	16.8	33.6	50.4	84.0	117.7	151.3
	$n_{WT}, \Delta = 0.40$	17.9	35.8	53.7	89.5	125.3	161.2
	n_{HP}	15.9	31.8	47.7	79.6	111.4	143.2
10	n_P	20.4	40.8	61.2	102.0	142.8	183.6
	n_B	16.3	32.6	48.9	81.5	114.2	146.8
	$n_{WT}, \Delta = 0.10$	16.5	33.1	49.6	82.6	115.7	148.7
	$n_{WT}, \Delta = 0.25$	17.1	34.2	51.2	85.4	119.5	153.7
	$n_{WT}, \Delta = 0.40$	18.4	36.8	55.3	92.1	129.0	165.8
	n_{HP}	16.2	32.4	48.6	81.0	113.4	145.8
15	n_P	21.0	42.0	62.9	104.9	146.9	188.8
	n_B	16.4	32.8	49.2	81.9	114.7	147.5
	$n_{WT}, \Delta = 0.10$	16.6	33.2	49.8	83.0	116.2	149.4
	$n_{WT}, \Delta = 0.25$	17.2	34.4	51.6	86.0	120.4	154.8
	$n_{WT}, \Delta = 0.40$	18.6	37.3	55.9	93.2	130.5	167.8
	n_{HP}	16.4	32.9	49.3	82.2	115.0	147.9
20	n_P	21.4	42.7	64.1	106.9	149.6	192.3
	n_B	16.4	32.8	49.3	82.1	114.9	147.8
	$n_{WT}, \Delta = 0.10$	16.7	33.3	50.0	83.3	116.6	149.9
	$n_{WT}, \Delta = 0.25$	17.3	34.5	51.8	86.3	120.8	155.4
	$n_{WT}, \Delta = 0.40$	18.8	37.5	56.3	93.8	131.4	168.9
	n_{HP}	16.6	33.3	49.9	83.2	116.5	149.7

K: Number of analysis, n: Sample sizes for tests; n_P: Pocock; n_B: O'Brien & Fleming; n_{WT}: Wang & Tsiatis; n_{HP}: Haybittle-Peto.

Table 10. Sample sizes for $\alpha = 0.05$, $(1 - \beta) = 0.80$, and $d = 1.5$.

K	n	$\sigma^2 = 1$	$\sigma^2 = 2$	$\sigma^2 = 3$	$\sigma^2 = 5$	$\sigma^2 = 7$	$\sigma^2 = 9$
1	n_f	7.0	13.9	20.9	34.8	48.8	62.7
5	n_P	8.6	17.1	25.7	42.8	60.0	77.1
	n_B	7.2	14.3	21.5	35.8	50.1	64.5
	$n_{WT}, \Delta = 0.10$	7.2	14.5	21.7	36.2	50.7	65.2
	$n_{WT}, \Delta = 0.25$	7.5	14.9	22.4	37.4	52.3	67.2
	$n_{WT}, \Delta = 0.40$	8.0	15.9	23.9	39.8	55.7	71.6
	n_{HP}	7.1	14.1	21.2	35.4	49.5	63.7
10	n_P	9.1	18.1	27.2	45.3	63.5	81.6
	n_B	7.2	14.5	21.7	36.2	50.7	65.2
	$n_{WT}, \Delta = 0.10$	7.3	14.7	22.0	36.7	51.4	66.1
	$n_{WT}, \Delta = 0.25$	7.6	15.2	22.8	37.9	53.1	68.3
	$n_{WT}, \Delta = 0.40$	8.2	16.4	24.6	40.9	57.3	73.7
	n_{HP}	7.2	14.4	21.6	36.0	50.4	64.8
15	n_P	9.3	18.6	28.0	46.6	65.3	83.9
	n_B	7.3	14.6	21.8	36.4	51.0	65.5
	$n_{WT}, \Delta = 0.10$	7.4	14.8	22.1	36.9	51.7	66.4
	$n_{WT}, \Delta = 0.25$	7.6	15.3	22.9	38.2	53.5	68.8
	$n_{WT}, \Delta = 0.40$	8.3	16.6	24.9	41.4	58.0	74.6
	n_{HP}	7.3	14.6	21.9	36.5	51.1	65.7
20	n_P	9.5	19.0	28.5	47.5	66.5	85.5
	n_B	7.3	14.6	21.9	36.5	51.1	65.7
	$n_{WT}, \Delta = 0.10$	7.4	14.8	22.2	37.0	51.8	66.6
	$n_{WT}, \Delta = 0.25$	7.7	15.3	23.0	38.4	53.7	69.1
	$n_{WT}, \Delta = 0.40$	8.3	16.7	25.0	41.7	58.4	75.1
	n_{HP}	7.4	14.8	22.2	37.0	51.8	66.5

K: Number of analysis, n: Sample sizes for tests; n_P: Pocock; n_B: O'Brien & Fleming; n_{WT}: Wang & Tsiatis; n_{HP}: Haybittle-Peto.

Pocock test required the largest sample size among the group sequential test types. O'Brien & Fleming and Haybittle-Peto tests had nearest sample sizes to fixed sample size design and order of these two tests changed with number of analyses. For example, when $K<15$ Haybittle-Peto test had the smallest sample size while O'Brien & Fleming test required the smallest sample size when $K\geq 15$. But Pocock test had always the largest sample size for all combinations. Wang & Tsiatis test was always placed between these two tests as for critical values. It had the same sample size with Pocock test when $\Delta=0.50$ and with O'Brien & Fleming test when $\Delta=0$. The sample size was close to O'Brien & Fleming test for small Δ values and started close to Pocock test with increasing Δ values. This condition was not varying according to number of interim analysis and other parameters related to calculation of sample size, as power, Type I error rate, effect size and variance. Sample sizes were almost similar for increasing effect sizes, and became same at $d=2.0$ and more effect sizes. It was mainly caused by the smallness of sample sizes such as 3.9. Because of the smallness of sample sizes, the differences between sample sizes for each test can not be observed and they were seen similar. So, advantage and disadvantage of each test in term of sample size can be compared for low effect sizes.

4. Discussion

Results obtained about sample size in all combinations almost same. Haybittle-Peto and O'Brien & Fleming tests have been required much smaller sample sizes comparing to other test types. Pocock test has been required the largest sample size for all combinations. Wang & Tsiatis test has been always required sample size that placed between O'Brien & Fleming and Pocock tests.

The reason for requiring small sample size of O'Brien & Fleming test comparing to Pocock test, can be understood from the formula for calculating the Z statistic, Equality (2.2.1). It can be seen that, effect of effect size on expected value of test statistic $E(Z_k)$ increases with number of analysis. So, the power of test achieves mainly later analyses [4,5].

Maximum sample size requiring in group sequential designs increases as the number of analyses increase. But, basic goal of group sequential designs is evaluating the advantage of early stopping through interim analysis [6,7, 13]. It takes into consideration that, maximum sample size is only required when there is no positive result in all interim analyses and the trial goes on to the final analyses.

Critical values of group sequential tests in interim analyses were ordered in a different manner, changing for each interim analysis. In addition, order of tests in term of critical values changing according to number of analysis (K). Pocock test had the lowest critical values for early interim analyses while O'Brien & Fleming test had the lowest critical values for latter analyses. Number of interim analysis in which this changing occurred varied according to number of analysis (K).

$C_B(K,\alpha)\cdot\sqrt{(K/k)}$ values increase because of increase in $\sqrt{(K/k)}$ values, and so the difference in critical values according to other tests in initial interim analyses increase as planned total number of analyses (K) increase. $\sqrt{(K/k)}$ values start to decrease from first analysis to final analyses and therefore $C_B(K,\alpha)\cdot\sqrt{(K/k)}$ values start to decrease, so the critical values according to other tests are lower in latter analyses. Similarly, Wang & Tsiatis test shows same manner for critical values. $C_{WT}(K,\alpha,\Delta)\cdot(K/k)^{\Delta-1/2}$ values increase because of increase in $(k/K)^{\Delta-1/2}$ values in initial interim analyses as planned total number of analyses (K) increase, and there is a big difference in terms of critical values according to other tests. It starts to decrease as the analyses goes on, and lower than other tests in latter analysis.

The number of analyses performed is important as the test type used. In some conditions, 1 or 2 interim analyses may be effective for decreasing sample size, and generally 4 or 5 interim analyses are sufficient. Accordingly, for a group sequential design using O'Brien Fleming test, $K=10$ interim analyses have been the optimum. In addition, for a group sequential design using Pocock test, $K>5$ interim analysis seems unreasonable.

As a result, these four test types have several advantages and disadvantages. In a group sequential trial, decision of test type using to analysis the trial data, based on a few criteria: 1) whether early termination is important or not; 2) reducing sample size; 3) the issue of trial; 4) whether reaching the subject easy or not; 5) detecting minimal effect sizes. In the conditions that, reaching subjects is hard or studying smaller sample size because of high risk, the test which provides that detect smaller treatment differences at the first interim analyses can be preferred.

REFERENCES

[1] D. L. DeMets, "Sequential Designs in Clinical Trials," *Cardiac Electrophysiology Review*, 1998, Vol. 2, No. 1, pp. 57-60. doi:10.1023/A:1009954810211

[2] S. C. Chow and J. P. Liu, "Design and Analysis of Clinical Trials: Concepts and Methodologies (Wiley Series in Probability and Statistics)," 2nd Edition, Wiley-Blackwell,

Hoboken, 2004.

[3] P. C. O'Brien, "Data and Safety Monitoring," In: P. Armitage and T. Colton, Eds., *Encyclopedia of Biostatistics*, Vol. 2, 2005, pp. 1362-1371.

[4] R. Aplenc, H. Zhao, T. R. Rebbeck and K. J. Propert, "Group Sequential Methods and Sample Size Savings in Biomarker-Disease Association Studies," *Genetics*, Vol. 163, 2003, pp. 1215-1219.

[5] C. Jennison and B. W. Turnbull, "Group Sequential Methods with Applications to Clinical Trials," Chapman & Hall/CRC, Boca Raton, 2000.

[6] M. Mazumdar and A. Liu, "Group Sequential Design for Comparative Diagnostic Accuracy Studies," *Statistics in Medicine*, Vol. 22, No. 5, 2003, pp. 727-739. doi:10.1002/sim.1386

[7] M. Mazumdar, "Group Sequential Design for Comparative Diagnostic Accuracy Studies: Implications and Guidelines for Practitioners," *Medical Decision Making*, Vol. 24, No. 5, 2004, pp. 525-533. doi:10.1177/0272989X04269240

[8] B. Tasdelen and E. A. Kanik, "The Formulae and Tables to Determine Sample Sizes for Classical Hypothesis Tests," *University of Mersin School of Medicine Medical Journal*, Vol. 4, 2004, pp. 438-446.

[9] S. C. Chow, J. Shao and H. Wang, "Sample Size Calculations in Clinical Research," Marcel Dekker, Inc., New York, 2003.

[10] J. M. Lachin, "Sample Size Determination," In: P. Armitage and T. Colton, Eds., *Encyclopedia of Biostatistics*, Vol. 7, 2005, pp. 4693-4704.

[11] E. Lakatos, "Sample Size Determination for Clinical Trials," In: P. Armitage and T. Colton, Eds., *Encyclopedia of Biostatistics*, Vol. 7, 2005, pp. 4704-4711.

[12] H. H. Muller and H. Schafer, "Adaptive Group Sequential Designs for Clinical Trials: Combining the Advantages of Adaptive and of Classical Group Sequential Approaches," *Biometrics*, Vol. 57, No. 3, 2001, pp. 886-891. doi:10.1111/j.0006-341X.2001.00886.x

[13] T. G. Karrison, D. Huo and R. Chappell, "A Group Sequential, Response-Adaptive Design for Randomized Clinical Trials," *Controlled Clinical Trials*, Vol. 24, No. 5, 2003, pp. 506-522. doi:10.1016/S0197-2456(03)00092-8

